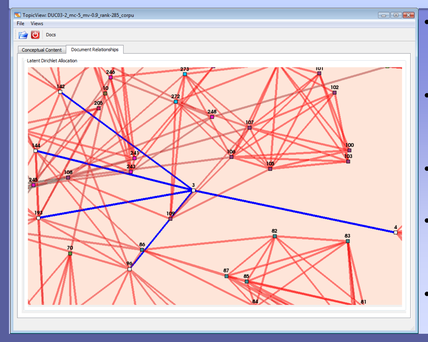
TopicView: Understanding Document Relationships Using Latent Dirichlet Allocation (LDA) Models

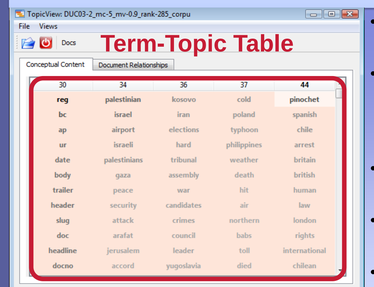
Patricia J. Crossno, Andrew T. Wilson, Daniel M. Dunlavy, and Timothy M. Shead (Sandia National Laboratories)

Unexpected Topic Connections

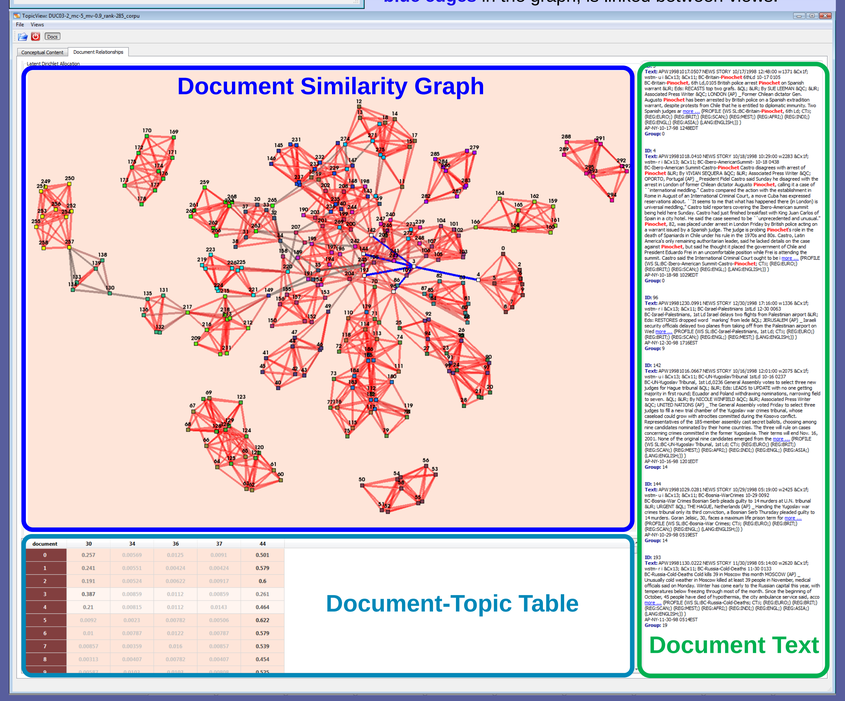


- LDA model of newswire articles from 2003 Document Understanding Conference (DUC) contest has unexpected links between articles.
- For example, documents **3** and **4** are articles on **Pinochet's arrest** in Britain (topic **44**), yet **3** is connected to articles on unrelated topics.
- Document **96** is an article describing Israel closing a **Palestinian airport** (topic **34**).
- Documents **142** and **144** are articles about electing judges to a **Yugoslav war crimes tribunal** in the Hague (topic **36**).
- Document **193** discusses **cold weather deaths** in Moscow (topic **37**).

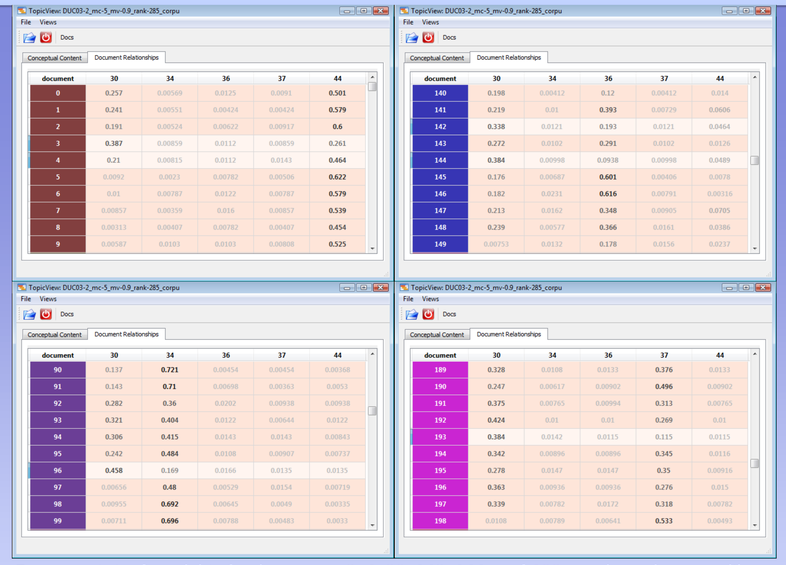
TopicView Model Exploration



- **Term-Topic Table** lists terms in order of importance within each topic. Term weight shown by **text darkness**.
- **Document Similarity Graph** displays document relationships. Edge weights color-coded red (high) to gray (low). Documents (nodes) colored by human generated topic groups provided in DUC contest data.
- **Document-Topic Table** provides document weights within each topic. Weight color-coded in **text darkness**.
- **Document Text** provides full text of selected documents with selected terms in Term-Topic Table shown in **red**.
- Document selection, highlighted by **white nodes** and **blue edges** in the graph, is linked between views.

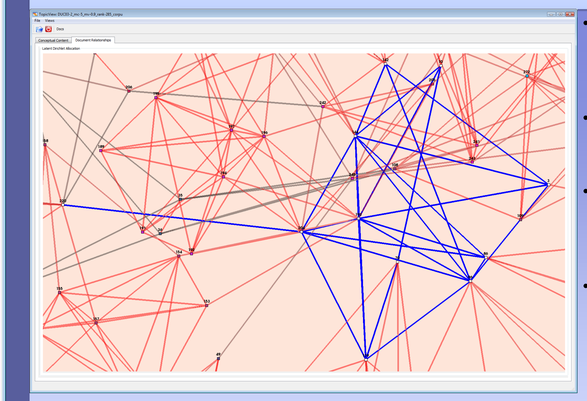


Document-Topic Weight Pattern



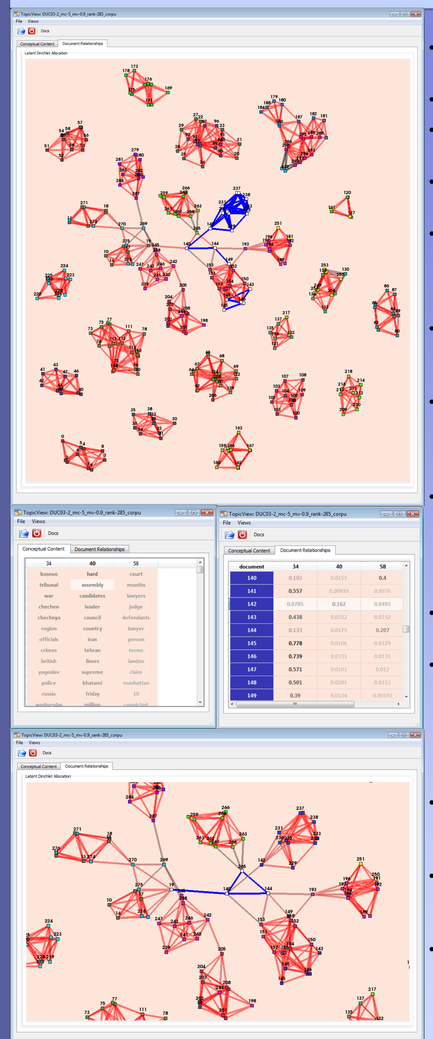
- **Document -topic** weights for documents **3**, **4**, **96**, **142**, **144**, **and** **193** are shown in rows with pale backgrounds for topics (columns) **30**, **34**, **36**, **37** and **44**.
- **Topic 30** is included because many of these documents are strongly weighted against it. Topic 30's most significant terms consist of XML header tags used in articles from the **Associated Press**. (Topic 31, not shown, similarly groups articles from the New York Times).
- **An interesting pattern** emerges (for all articles except 4, which is connected to 3 by concept) – the **articles are more strongly weighted in their source topic** than in the topic representing their conceptual content.
- **Hypothesis: Documents matching this pattern are the source of many of the edges between disjoint topics.**

Bridging Documents



- We test our hypothesis by listing all documents whose strongest weights are AP or NYT topics.
- For each document in the list, we check all linked documents in the graph (like we did for 3).
- If linked documents share conceptual content (like 4), then the document fails the hypothesis.
- Of 297 documents, 33 fit the hypothesis. Of these, only 11 survived the test, having links not supported by conceptual content (including **3**, **96**, **142**, **144**, and **193**) .
- These 11 are in the center of the graph and tend to link with one other, impacting layouts of their associated clusters. We call them **bridging documents**. They are **all short AP articles**, where the terms in the headers outweigh the text in the news content (NYT articles tend to be longer).
- **Revised hypothesis: Only documents whose conceptual content is outweighed by the source content will display this bridging pattern that links disjoint topics.**

Merged Topics & Bridging Topics



- We test our revised hypothesis by rerunning LDA on **headlines and story bodies only**.
- **Many edges linking disjoint topics disappear**.
- Document clusters separate and layout improves.
- **AP & NYT topics disappear**, freeing 2 topics. DUC data consists of 30 labeled topic groups.
- Previous topic **36** splits into new topics **34** & **40**, separating **Yugoslav war crimes tribunal** and **Iranian election** stories (highlighted by **white nodes** and **blue edges** in the upper graph).
- Now articles on **Chechen kidnappings** have been **merged** with **Yugoslav war crimes tribunal** articles in topic **34.**
- Document **142** still links topics **34** and **40.** As with the AP source, **142** is more strongly linked to topic **40** (**Iranian elections**) than to its conceptual topic, **34** (**tribunal-kidnappings**).
- In topic **40, *assembly*** and ***candidates*** are the 2nd and 3rd strongest terms. In document **142, *assembly*** and ***candidates*** appear 3 and 4 times each. These terms dominate the other terms and strongly link **142** to topic **40**.
- Document **142** remains a **bridging document**, though the conceptual link is valid here.
- Narrowly-focused topics appear, i.e. topic **58** linking trial-related stories (shown in the lower graph linked by **blue edges**, documents **19**, **140**, **144**, and **256**). This **bridging topic (**similar to the AP topic) connects documents more strongly than their conceptual groups.
- **Choice of topic count impacts clustering,** given merging and splitting of topics. Tried counts from 28 to 78. Correct count unclear.
- More topics will not necessarily separate combined topics or reduce edges. Topics can merge and bridging topics appear.
- With more topics, document weights within some topics may become so low that they appear to be **noise**.

Contact Information:
Patricia J. Crossno: pjcross@sandia.gov
Andrew T. Wilson: atwilso@sandia.gov
Daniel M. Dunlavy: dmdunla@sandia.gov
Timothy M. Shead: tshead@sandia.gov

Acknowledgements:
This work was funded by the Laboratory Directed Research & Development (LDRD) program at Sandia National Laboratories.